Docket No. 1751.1001

TITLE OF THE INVENTION

A SYSTEM AND METHOD FOR UNIQUELY IDENTIFYING PERSONS

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The present invention is directed to the field of data searching. More particularly, the present invention relates to uniquely identifying a person when only minimal information about a person is available, such as a name and age, or a name and address, as located in one database source, and comparing that data in a separate source which has different datasets to thereby match one against the other.

2. Description of the Related Art

[0002] In April 2003, to restore investor confidence, various brokerage and finance firms agreed with state and federal regulators to comply with a Voluntary Initial Public Offering (IPO) Agreement. The Agreement may be found at www.sec.gov/news/press/globalvolinit.htm (as of October, 2003). Under the Agreement, participating firms agreed to implement reasonable procedures to ensure that they do not allocate "hot" IPO securities to the accounts of officers and directors of qualified publicly traded companies. The firms also agreed not to allocate such securities to the accounts of immediate family members of officers and directors of publicly traded companies.

[0003] However, it is difficult to determine whether any person, including an account holder, is an officer or director of a publicly traded company. There is no listing of uniquely identified (e.g. by social security number) officers and directors. In the past, determining whether a name and age or name and address correspond to a particular individual has required manual investigation. What is needed is a system and method that will allow participating firms to automatically identify, with reasonable certainty, an account holder or customer, or immediate family member thereof, as an officer or director of a publicly traded company based on non-uniquely identified names of such officers and directors and based on information of the

population at-large of which the account holder is a member.

SUMMARY OF THE INVENTION

[0004] It is an aspect of the present invention to provide a system and method to determine with reasonable certainty the true identify of a non-uniquely identified name and age or name and address.

[0005] It is another aspect of the present invention to provide a system to automatically determine which accounts of a firm are held by an officer or director of a publicly traded company.

[0006] It is yet another aspect of the present invention to combine various disparate sources of public records into a combined public records dataset, and to use the public records dataset to help uniquely identify an individual corresponding to a non-unique name, or to identify immediate family members or cohabitants corresponding to the non-unique name.

[0007] It is still another aspect of the present invention to combine various sets of Security and Exchange Commission (SEC) records to obtain a list of non-uniquely identified names, and one or more of an associated address and age.

[0008] It is a further aspect of the present invention to determine whether a named individual customer of a firm corresponds to a record of an officer or director of a publicly traded company based on a measure of how unique the individual customer's name is.

[0009] It is another aspect of the present invention to match a name and age/address with a uniquely identified individual, when the name does not have an associated identifier or other indicia of uniqueness such as a social security number.

[0010] It is yet another aspect of the present invention to provide a system that combines records of the public at large to find sets of addresses of uniquely identified persons, and which

determines whether a person is related to an officer or director of a publicly traded company by referring to the sets of common historical addresses.

[0011] The above aspects can be attained by a system and method that determines whether a non-uniquely identified name substantially corresponds to a uniquely identified person. A source dataset of uniquely identified persons is accessed, where the source dataset has records including, for each uniquely identified person, a source name, a source unique identifier, a source date of birth, and a source address. A target dataset of non-uniquely identified persons is also accessed, where the target dataset has records that include, for each non-uniquely identified person, a target name, and either (1) a target age and a target age-date indicating an exact or approximate date of which the target age was recorded, or (2) a target address. For a particular source person in the source dataset, whether the particular source person corresponds to a particular target person in the target dataset is determined automatically in accordance with the accessing.

[0012] These together with other aspects and advantages which will be subsequently apparent, reside in the details of construction and operation as more fully hereinafter described and claimed, reference being had to the accompanying drawings forming a part hereof, wherein like numerals refer to like parts throughout.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] Figure 1 shows a generic system that may be used for matching uniquely identified account holders (firm customers) with weakly-identified names of officers and directors of publicly traded companies.

[0014] Figure 2 shows an overall process for matching all account holders or customers when some officers and directors are uniquely identified.

[0015] Figure 3 shows aggregated datasets.

[0016] Figures 4A-4F show tables/files of insider trading information and business records 100, 104, 108, 112, 116, and 120 that are preferably used as the SEC data sources 58.

[0017] Figure 5 shows a process for processing a list of well-identified names against weakly or non-uniquely identified names.

[0018] Figure 6 shows an example of address matching.

[0019] Figure 7 shows an example of name-uniqueness matching 146.

[0020] Figure 8 shows one of many possible hardware configurations that may be used to implement embodiments of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

OVERVIEW: NEED TO IDENTIFY RESTRICTED TRADERS

[0021] As discussed in the "Background of the Invention", it is difficult to identify whether a well-identified person is an officer or director of a publicly traded company. Put another way, participating firms are charged with the difficult task of knowing whether the "Chris Smith" associated with a particular well-identified account is the same "Chris Smith" who is an officer or director of a publicly traded company, when there may be many people named "Chris Smith" in the at-large population. Despite significant need to identify such people, automated identification of an officer or director has not previously been accomplished.

[0022] A firm participating in the IPO Agreement mentioned in the "Background of the Invention" manages investment accounts for their customers. An account may have securities owned by the account holder. The firm's customer, who is the holder of the account, may be well identified to the firm. For example, for each account, the firm may have the account holder's social security number or equivalent. This is a unique number, which, if accurate, uniquely identifies the real-world persona of the holder of the account. Although an investment

firm may have a high level of confidence that the identity of an account holder is correct, that unique identity is difficult to match to a bare list of names, such as of officers and directors, with only weak or non-unique associated identity information such as age, address, etc. Furthermore, such a list of named officers and directors is not readily available, and must be pieced together using data synthesis tools and algorithms that sift through hundreds of databases to come up with a match.

[0023] Figure 1 shows a system for matching uniquely identified account holders (firm customers) with weakly-identified names of officers and directors of publicly traded companies. In general, a dataset of unidentified and potentially non-unique names 20 (e.g. names of officers and directors), a dataset of known persons 22 (e.g. firm customer list), and a dataset of public records 24 are provided for access 26. Then, preferably using a multi-stage matching or elimination process, non-unique names in the non-unique name dataset 20 are matched with or eliminated using unique names in the dataset of known persons 22 and using the public records 24 to help identify matches 30.

[0024] Figure 2 shows an overall process for matching all account holders or customers when some officers and directors are uniquely identified, as for example by SSN. As seen in Figure 3 (discussed later), approximately 35 % of the relevant SEC records provide a related social security number (SSN), however, such information comes from an information data provider that has not verified the correctness of the data being reported to them on SEC forms 3, 4, 5 and 144. Typically, the data provider will not have verified that the self-confessed information placed on the form is correct, and the data provider will not have verified that the SSN is for the exact name of the person that it has been assigned to by the SSA and many other variables. Therefore, in an overall process for determining which account holders are officers or directors, an initial step, after acquiring 50 the datasets 20, 22, and 24, is to search 52 for the high-certainty matches. That is to say, names of officers and directors in dataset 20 that have an associated SSN are matched with records in the uniquely identified dataset 22 that have a matching name and SSN (or only a matching SSN). The overall process continues with a search 54 to find matches where SSN is not available, by using the dataset of public records 24 to link the non-unique names with the unique names. Finally, the names and identifiers may

be crosschecked 56 for further assurance.

PREPARING AND ACQUIRING THE DATA

[0025] Figure 3 shows aggregated datasets. The public records dataset 24/50 is preferably the first dataset that is obtained or accessed. In a preferred embodiment, a set of 23 public record data sources 52 are combined into one public records network (PRN) dataset 50. The public data sources 52 are referred to as "public" because they generally contain records of information of a public provenance or of publicly conducted transactions. The public data sources 52 may in practice be public, private, proprietary, or restricted-access databases. Although most of the data sources 52 can be obtained commercially, some data sources 52 can only be obtained en masse under certain legal conditions (restricted access). The PRN dataset 50 will contain many instances of like-named people within the subject at-large population. This is one reason why name matching is difficult; the name of a weakly identified officer or director can potentially correspond to one of many different like-named individuals in the at-large population. One skilled in the art will appreciate the similarity of the records of some individuals in a statistically large population (e.g. millions of people).

[0026] Examples of possible data sources 52 include, but not are not limited to: 1 - data of birth, 2 - driver's license, including name and address, 3 - alias or also-known-as names, 4 – other SSNs, 5 - other names associated with an SSN, 6 - addresses associated with a subject, 7 - real property ownership, 8 - deed transfers, 9 - vehicles registered at subjects' addresses, 10 - watercraft, 11 - FAA aircraft registration, 12 - UCC filings, 13 - bankruptcies, liens, and judgments, 14 - professional licenses, 15 - FAA pilot licenses, 16 - DEA controlled substance license, 17 - business affiliations, 18 - relatives of other people who have the same address as the subject, 19 - licensed drivers at subject's address, 20 - neighborhood phone listings for subject's addresses, 21 - banking, financial, and credit relationships, 22 - credit report data, that is restricted under FCRA, 23 - asset-based records. Of these 23 public data sources 52, sources 1, 3, 4, 5, 6, and 18 are the most significant for the present invention. The public data sources 52 are combined. The public data sources 52 or the combined PRN dataset 50 may be

commercially obtained from Thomson Analytics. It is important that the PRN dataset 50 contain names, and where available, SSNs, dates of birth (d.o.b.), and addresses. Typically, by mining and piecing together the public data sources 52, across the records, it is possible to have the SSN for 95% of the subjects, the d.o.b. for 50% of the subjects, and the address for 70% of the subjects. The PRN dataset 50 may contain approximately 20 billion records. There is an assumption that the subjects or names to be matched are within the general population corresponding to the PRN dataset 50, that is to say, the mass of persons whose information is found in the PRN dataset 50.

[0027] Preferably, the PRN dataset 50 is used to cleanse 54 the Customer Profile Source (CPS) dataset 56, although other data sources or only algorithms may be used for cleansing 54. Cleansing 54 can involve any number of well-known techniques, including spelling correction, comparison for consistency with public records carrying the same information (e.g. d.o.b.), and so on. Although substantially all CPS dataset 56 records will be populated with a name, SSN, d.o.b., and address, the records are preferably cleansed to improve their accuracy. The SSN and d.o.b. are verified and updated if necessary. All past addresses of the subject ore obtained for the purpose of later checking to obtain the names and identities of spouses or minor children A cleanse code, discussed in the Appendix, can be added to CPS dataset 56 records to indicate a level of quality or reliability of each record.

[0028] A weakly identified SEC dataset 57 is obtained by combining various SEC data sources 58, including insider trading information, SEC Form filings, and the like. The SSN will be available in 35% of all cases. The age will be available in 70% of cases. The address will be available in 100% of cases, however the address can, without indication, correspond to a work location, a residential location, and can be either a present or past location of the subject. Preferably, the existence of a record itself is used as the information that indicates that a record's named subject is or was an officer or director of a publicly traded company. As discussed later, the information gaps in the SEC dataset 57 are addressed by using different matching techniques according to the information available for a given subject. The SEC dataset 57 contains records relating to approximately 500,000 individuals. Again, the individuals in the SEC dataset 57 are assumed to be from among the same general population

that corresponds to the PRN dataset 50 and the CPS dataset 56. One skilled in the art will appreciate that a population refers to most people inhabiting one or more countries, regions, commonly governed areas, etc.

[0029] Figures 4A-4F show tables/files of insider trading information and business records 100, 104, 108, 112, 116, and 120 that are preferably used as the SEC data sources 58. Such tables are commercially available. The header file 100 shown in Figure 4A has information included in the header on SEC Forms 3, 4, 5, and 144. The header information can be linked to the transactional files through the Document Control Number (DCN). The header file 100 also captures insider filings with header information only, which is typical of the SEC Form 3. The records of the header file 100 generally span from January 1986 to the present. The header file 100 is the primary indicator of whether a Form's subject was serving in the role of an officer or director.

[0030] The table one file 104 shown in Figure 4B contains most transaction and holdings information filed on SEC Forms 3, 4, and 5. The table one file 104 has several value-added fields including a cleanse indicator that identifies whether the data was cleansed using external data sources, and an indicator of the degree of confidence in each data record. Cleanse indicator codes are described in the Appendix. The records of the table one file 104 also span from January 1986 to the present. Cleansing services may be commercially purchased.

[0031] The table two file 108 shown in FIGURE 4C contains most transaction and holdings information filed on SEC Forms 3, 4, and 5. The data in the table two file 108 includes open market derivative transactions as well as information on the award, exercise, and expiration of stock options. The records of the table two file 108 generally span from January 1996 to the present.

[0032] The Form 144 proposed sale file 112 shown in Figure 4D is derived from SEC Form 144 filings. This data includes the expected date of sale of securities, the number of securities to be sold, the estimated market value of the proposed sale, and the name of the executing broker. The records of the Form 144 proposed sale file 112 span from June 1996 to the present.

[0033] The individual returns file 116 shown in Figure 4E is derived from SEC Form 144 filings and includes the expected date of sale, the number of securities to be sold, the estimated market value of the proposed sales, and the name of the executing broker. The records of the individual returns file 116 span from June 1996 to the present.

[0034] The company information file 120 shown in Figure 4F provides company specific identifiers including security ID, ticker, company name, sector, and industry. The security ID is the link back to the insider transactions files 100, 104, 108, and the form 144 proposed sale file 112. The records in the company information file span from June 1986 to the present.

[0035] Given the datasets 50, 56, and 57 discussed above, it is possible to perform the matching methods discussed below.

MATCHING KNOWN PERSONS TO RECORDS OF NON-UNIQUELY IDENTIFIED RECORDS

[0036] As discussed above, a purpose of the present invention relates to matching a loosely identified person/name to a well-identified person/name. In the application of identifying former officers or directors for security trading firms, it is noted that because under the Hot IPO Agreement a participating firm need only "reasonably" identify whether an account holder is a restricted trader, it is not necessary to find matches with high certainty. Rather, finding a match that has only a reasonable probability (say 50%) of being correct will satisfy a firm's obligation.

[0037] Figure 5 shows a process for matching a list of well-identified names against weakly or non-uniquely identified names. Initially, where a strong link is available, easy matches are found by matching 140 those SEC records for which an SSN is available, by comparing, in the case of weakly identified SEC officers and directors, SEC names and SSNs to CPS names and SSNs. The accuracy of an SSN match 140 is improved by the initial cleansing 54 of the CPS, dataset 56 and by cleansing of the SEC dataset 57.

[0038] For those SEC records where a match 140 by SSN or some other identification key is not possible, a name and age match 142 is performed. Generally, SEC records have a date that indicates when the record was created or a point in time when the data of the record was

obtained, for example by the filing of an SEC Form 144. When an SEC name matches a CPS name, it is possible to determine whether the two names correspond to the same individual by comparing the date-adjusted SEC age (or an equivalent d.o.b) with the CPS d.o.b/age. A match 142 will indicate that the SEC record and CPS record with the same name are reasonably likely to correspond to the same individual.

[0039] For those SEC records of SEC dataset 57 where an SSN and age/date are not available, a match 144 based on name and address(es) is used. Where the SEC address and the CPS address for the same name match, then a match 144 is assumed. Furthermore, using address records from the PRN dataset 50, it is also possible to determine a match where the SEC address and the CPS address are not the same. In this case, a set of historical addresses from the PRN dataset 50, preferably going back 15 years, are linked to the well-identified CPS subject. The historical address(es) preferably include all known work or residential addresses of the subject. If the SEC address matches one of the historical addresses, then a match 144 is assumed.

[0040] In cases where only an SEC name is available, name-uniqueness matching 146 is used to determine a reasonable match. See the discussion of Figure 7.

[0041] Finally, for the application of determining whether a subject is an officer or director of a publicly traded company, the obligation of a firm to identify close relatives of an officer or director can be met by using cohabitation as an indicator of familial or relational immediacy. The PRN address information is used to find 148 people who have co resided with an officer or director of a publicly traded company. The algorithm is similar to that discussed in step 144. Preferably, a determination of whether a person is an immediate family member is based on whether that person shared an address with the officer or director for 5 years (or some other period), or for two or more consecutive addresses. For example, if CEO Chris Smith resided at address1 for 5 years, and Pat Smith also resided at address1 for the same period of time, then Pat Smith is assumed to be closely related to Chris Smith. Or, if the Chris Smith resided at address1 and then address2, and if Path Smith also resided at address1 and then address2, then Pat Smith would also be assumed to be closely related Chris Smith. Potential close relations can be

derived from a number of sources, including the PRN dataset 50, the CPS dataset 56, and so on. Preferably, possible close relations are extracted from free-form text fields in the CPS dataset 56 records, which may contain ad-hoc information related to an account holder, such as trust or inheritance information.

[0042] Although steps 142, 144, and 146 are shown in sequence in Figure 5, these matching algorithms may be performed in different orders, in different, combinations, and so on. For example, steps 142, 144, and 146 may all be used when all of the non-SSN SEC information is available (e.g. name, age, and, address). The results may be used to return all possible name iterations, ranked in order of likelihood of identity. For example, when an SEC name's age and address are both available, and both match a particular CPS record, the likelihood of true identity between them will be higher than if the address was not available and a match was determined based only on name and age.

[0043] Figure 6 shows an example of address matching technique. In the address matching 144 discussed with reference to Figure 5, where an SEC record 160 matches the name of a CPS record 162, or optionally where an age/d.o.b. match also occurs, a set of addresses 164 from the PRN dataset 50 is used to match the addresses of the two records. The set of addresses 164 preferably includes all work or residential addresses from the previous 15 years that are associated with the Chris Smith of CPS record 162. Thus, whether the address of the SEC record 160 is or was a work or residential address, an address match can still be determined. In sum, identity can be established between the two Chris Smith records 160, 162 by determining that a same "Chris Smith" used two addresses (e.g. addr1 and addr3) held by one known person, each associated with one of the Chris Smith records 160, 162.

[0044] Figure 7 shows an example of name-uniqueness matching 146. The name-uniqueness matching may be performed individually to confirm a name match, or it may be performed in combination with the matching of other available pieces of information. In the example shown in Figure 7, a surname "xaxy" is matched to a pre-existing list of name uniqueness rankings. In this example, the uncommon name of "xaxy" has a ranking of .999, which is used to determine that the SEC record 160 and the CPS record 162 matches with

reasonable certainty. The uniqueness of a last name may also be taken into account, or a combined surname and last name uniqueness rating may be used.

[0045] In test cases processed according to the above, success rates of approximately 50% have been achieved. Adjusting some parameters such as the 15 year mark for addresses, the 5 year residency parameter, changing the cut-off point for uniqueness of a name, and so on, may all be altered according to a desired balance between accuracy and inclusiveness.

[0046] Figure 8 shows one of many possible hardware configurations that may be used to implement embodiments of the present invention. Generally, information such public records mentioned above may be transmitted from servers 170 of a data provider such as Thomson Analytics, over a network 172, to servers 174 implementing aspects of the invention mentioned above. The servers 172 and 174 may include one or more preferably commercial databases 176. The information needed by servers 172 may be provided by servers 174 either wholesale where it is then searched at servers 172, or it may be provided be maintained and searched at servers 172 as needed. Searches using aspects of the present invention may be conducted by a user using a workstation 178, which may include a processing unit 180, a display 182, and input devices 184. The workstation 178 may function as a client accessing the servers 174 through the network 172, for example using HTTP or other IP-based protocols. Not shown are other computers, for example SEC servers or servers of trading firms that may provide the SEC and account holder information discussed above. Batch exchanges of data, and updates to the trading firms over the network 172 (based on search results) may also be conducted.

OTHER APPLICATIONS

[0047] The methods discussed above are not limited to the application of identifying officers and directors of publicly traded companies. The method of linking weakly identified names with strongly identified names based on common address, age/d.o.b, name uniqueness, etc. can be extended to other applications. For example, aspects of the invention may be used to satisfy duties imposed by the Patriot Act and the Know Your Customer Act.

[0048] Aspects of the present invention have been described with respect to a system and method that determines whether a non-uniquely identified name substantially corresponds to a uniquely identified person. A source dataset of uniquely identified persons is accessed, where the source dataset has records including, for each uniquely identified person, a source name, a source unique identifier, a source date of birth, and a source address. A target dataset of non-uniquely identified persons is also accessed, where the target dataset has records that include, for each non-uniquely identified person, a target name, and either (1) a target age and a target age-date indicating an exact or approximate date of the target age, or (2) a target address. For a particular source person in the source dataset, whether the particular source person corresponds to a particular target person in the target dataset is determined automatically in accordance with the accessing.

[0049] In a preferred embodiment the results of the inquiry are automatically compared against the profile of the client, which is updated and sent back to a requestor in an encrypted format. A typical embodiment will be capable of performing 20,000 or more searches per day, and will return the clean data sets to a customer. It is also preferable to automatically validate certain fields of data as contained within the customer profile, such as the customer's True Name, True DOB, True Age, True Social Security account number, True current home address, True home phone number, True name of current spouse, and True maiden name or second name of spouse. Any anomalies are preferably highlighted in a NOTES section of the customer's profile.

[0050] The many features and advantages of the invention are apparent from the detailed specification and, thus, it is intended by the appended claims to cover all such features and advantages of the invention that fall within the true spirit and scope of the invention. Further, since numerous modifications and changes will readily occur to those skilled in the art, it is not desired to limit the invention to the exact construction and operation illustrated and described, and accordingly all suitable modifications and equivalents may be resorted to, falling within the scope of the invention.